The 'Celiac Paradox': Investigating Evolutionary Patterns and Selective Mechanisms of Genetic Risk Factors for Celiac Disease

Angeni Bai

Abstract

Celiac disease (CD) is an autoimmune disease which causes mild to severe gastrointestinal symptoms in its patients when gluten is consumed. The higher incidence of CD observed in regions with a longer history of gluten-containing cereal agriculture is known as the 'CD evolutionary paradox' (Singh et al. 2018). Further, previous studies have found a link between a longer history of wheat consumption and a higher frequency of the CD-predisposing *HLA* haplotype between countries, hypothesising that the haplotype was selected to protect against a pathogen related to tooth decay (Lionetti and Catassi 2014). Using purposely constructed computer algorithms, this study investigated whether this relationship was unique to CD risk-alleles and replicable across regions of genetic ancestry rather than simply countries (where ancestry is mixed). Global populations were grouped into eight ancestral regions and correlations between CD prevalence in these regions and allele frequency, as well as allele frequency and duration of wheat and rye agriculture (WRA) were investigated. Computational methods conducted linear regressions against 3 *HLA* risk-haplotypes, 41 background risk-alleles and 652 alleles selected randomly across the human genome. The results confirmed the CD evolutionary paradox, and revealed that WRA duration was associated not with *HLA* risk-haplotypes, but instead with a particular SNP, rs4686484, on the *LPP* gene, which plays a role in maintaining the lining of the small intestine damaged by CD. Thus, a novel hypothesis is proposed here that as CD prevalence increased alongside the adoption of a gluten-containing diet, the *LPP* gene experienced selection as a protective factor which counteracted the decrease in evolutionary fitness of affected individuals.

Definitions and Acronyms

[1]**1000 Genomes Project (1KGP)** - the largest public catalogue of human variation and genotype data.

[2]**Allele Frequencies Net Database (AFND)** - a publicly accessible database containing data on frequencies of alleles related to immune response collected from peer-reviewed studies

[3]**Allele** - a variant form of a gene at a given location on the chromosome, which may affect expression of the gene. In this report it is used interchangeably to denote both haplotype and SNP variants.

[4]**Ancestral region** - a geographical region with a shared genetic history, as identified by The Genographic Project.

[5]**Autosome** - one of the 22 non-sex chromosomes

[6]**Biopsy** - analysis of a tissue sample taken from the body to diagnose the presence of a disease

[7]**Celiac disease (CD)** - an autoimmune disease triggered by gluten

[8]**Haplotype** - a group of alleles that are inherited together

[9]**Human leukocyte antigen (*HLA*)** - a gene complex on chromosome 6 which is involved in the regulation of the immune system. It is particularly involved in CD.

[10]**Linkage disequilibrium** - when two or more alleles or genes are associated with one another non-randomly

[11]**Non-coding variant** - a variant which doesn't specifically provide information for the formation of proteins.

[12]**Phenotype** - an observable physical property of an individual which manifests due to their DNA.

[13]**Risk-allele/risk-haplotype/risk-SNP** - An allele/haplotype/SNP which increases the risk of developing a phenotype in individuals which carry it

[14]**Serological test** - an analysis of a blood sample for the presence of antibodies

[15]**Single nucleotide polymorphisms (SNPs)** - mutations at a single base pair in the genome that are present in more than 1% of the population.

[16]**Wheat and rye agriculture (WRA)** - agriculture of the two major gluten-containing cereals

Where clarity is required, terms and acronyms in the report are labelled with the corresponding reference number in superscript, eg. ...rs13132308 is another non-coding variant[11].

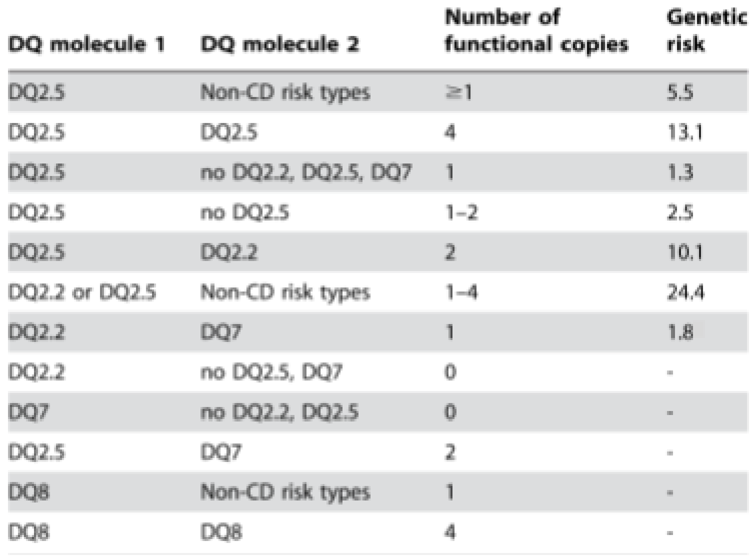
Additional note: Use of R^2 in the report should be interpreted as R-square values, not a reference to Allele Frequencies Net Database

Literature Review

Celiac disease (CD) is an autoimmune disease triggered by gluten, causing injury in the lining of the small intestine and characterised by mild to severe gastrointestinal symptoms (including diarrhea, malabsorption, abdominal pain and distension, bloating, vomiting, and weight loss) (Taylor et al. 2019).

The current medically accepted method of diagnosing CD is through a combination of serological[14] and biopsy[6] testing. This involves a blood sample analysis for the presence of elevated levels of particular antibodies (serum tissue transglutaminase IgA, anti-deamidated gliadin-related peptide IgA or IgG, endomysial antibody IgA), followed by an analysis of tissue samples collected from the bowel (Taylor et al. 2019).

The development of CD within an individual is a result of both genetic and environmental factors. The primary environmental trigger is the presence of wheat gluten and related proteins in the diet (Kagnoff 2007). Approximately 40% of genetic risk can be attributed to the presence of sets of DNA variations inherited together (haplotypes) on the *Human Leukocyte Antigen* (*HLA)* gene complex (Sams and Hawks 2013). DNA variations are referred to as alleles, and each individual carries two alleles for every gene. Fig. 1 details combinations of the molecules coded by their corresponding haplotypes[8], and their influence in increasing genetic risk. Up to 95% of celiacs are DQ2 positive (with about 90% being DQ2.5 positive) and the remaining 5% are DQ8 positive (Volta and Villanacci 2011). Hence, the presence of one of these haplotypes in an individual is necessary to develop CD.



DQ molecule 1	DQ molecule 2	Number of functional copies	Genetic risk
DQ2.5	Non-CD risk types	≥1	5.5
DQ2.5	DQ2.5	4	13.1
DQ2.5	no DQ2.2, DQ2.5, DQ7	1	1.3
DQ2.5	no DQ2.5	1–2	2.5
DQ2.5	DQ2.2	2	10.1
DQ2.2 or DQ2.5	Non-CD risk types	1–4	24.4
DQ2.2	DQ7	1	1.8
DQ2.2	no DQ2.5, DQ7	0	-
DQ7	no DQ2.2, DQ2.5	0	-
DQ2.5	DQ7	2	-
DQ8	Non-CD risk types	1	-
DQ8	DQ8	4	-

Fig. 1: Combinations of DQ molecules and their associated genetic risk for CD (Monsuur et al. 2008). Genetic risk for some combinations are undetermined.

Over 50 non-*HLA* mutations at single base-pairs in the DNA (single nucleotide polymorphisms, or SNPs) have been identified to predispose CD (Sams and Hawks 2014). They form a 'background risk network' that contributes up to 14% of genetic risk (Trynka et al. 2011).

The term ‘allele’ refers to a DNA variation in the genome, and in this report it is used interchangeably to refer to both haplotype and SNP variations.

Gluten is a set of proteins found in certain cereal grains such as wheat, oats, barley and rye. The agriculture of gluten-containing cereals began over 8000 years ago in a region including parts of the Middle East and Mediterranean Basin (Fig. 2), known as the Fertile Crescent (Curtis 2002). Adopting a gluten-free diet is the only known treatment for CD, which was recognised in the 1950s. Before this, the condition would have reduced reproductive fitness in its sufferers, leading to malnutrition, and even death in juveniles. Yet, the most recent meta-analysis of the global prevalence of CD found prevalence to be 0.7% (Singh et al. 2018), making it one of the most common food intolerance related disorders.

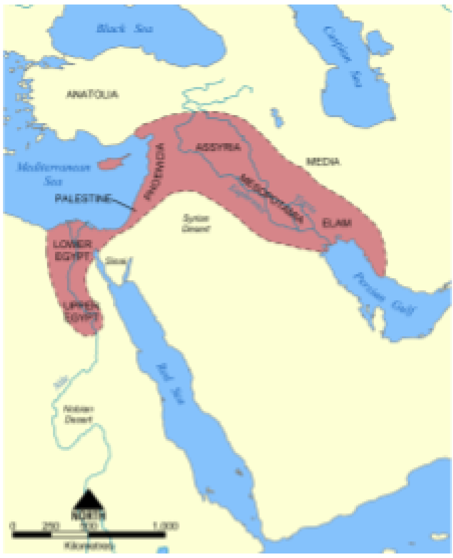


Fig. 2: Map showing the Fertile Crescent

Simoons (1978) hypothesised that the origination of wheat agriculture in the Fertile Crescent exerted negative selective pressure on the genes that predisposed CD. According to the theory of natural selection, individuals carrying genetic traits that decrease their reproductive fitness are less likely to reproduce and pass down those genes. Hence, the expectation should be to find a lower frequency of CD risk-alleles[13] and CD prevalence in geographic regions with a longer history of gluten-containing cereal agriculture. However, recent global reviews have found similar or higher prevalence of CD in the Middle East and Mediterranean when compared to other regions (Singh et al. 2018). This contradiction between evolutionary predictions and observed global patterns of CD prevalence is known as the ‘celiac disease evolutionary paradox’ or ‘celiac paradox’. Further, by analysis of global data, Lionetti and Catassi (2014) found evidence that a higher frequency of CD-predisposing haplotype HLA-DQ2 was correlated with a longer history of wheat consumption.

There are a number of unconfirmed hypotheses for this paradox (Sams and Hawks 2014). Two commonly investigated are:

1. If most genetic variations contributing to CD risk only have minor effects, the consequence of selection against CD on individual alleles[3] would be miniscule.
2. The same variants which increase CD risk may protect against other conditions and pathogens. Positive selection for non-CD phenotypes[12], especially those brought into significance by the agricultural revolution, would result in the paradox.

There have been significant limitations to the existing research investigating the celiac paradox. Global reviews such as Singh et al. (2018) investigated data by country, within which there is sometimes a high level of ethnic variation that isn't reflective of the spread of cereal agriculture. For example, geographically, the American region was introduced to gluten-containing cereals after 0 AD (Liu et al. 2019). Yet, 72.4% of the population are Europeans, Middle Easterners or North Africans (Humes, Jones & Ramirez 2011) who have genetic backgrounds that reflect the adoption of a gluten-containing diet between 5000 and 2500 BCE.

This problem also occurred in Lionetti and Catassi's (2014) comparison of *HLA*[9] risk-haplotype[13] frequency, CD prevalence, current wheat consumption, and duration of wheat consumption across countries. Their Australian haplotype frequencies derived from Indigenous individuals, while Australian studies on CD prevalence were performed in individuals of predominantly European origin. Since Indigenous Australians were introduced to gluten-containing cereals in the last two centuries, there is a discrepancy in genetic backgrounds that prevents the analysis from accurately representing the effect of wheat consumption history. In order to make conclusions about evolutionary history and the effect of the development of wheat agriculture, it would be more valid to categorise populations by their ancestry.

Another missing component in Lionetti and Catassi's work (2014) was an analysis of the relationship between each of the variables (ie. CD prevalence, HLA risk-haplotype frequency, etc.) and alleles which do not predispose for CD. Due to the random process known as genetic drift, allele frequencies within populations may change by chance alone, leading to differences between separate populations (Masel 2011). Without knowing the patterns of association that are due to genetic drift, it is impossible to determine if the significant correlation between risk-haplotype frequency and duration of wheat consumption is in fact attributable to natural selection.

Finally, Lionetti and Catassi (2014) used only European countries to investigate the association between duration of wheat consumption and risk-haplotype frequency. This was because the work detailing the history of wheat agriculture available at the time (Ammerman and Cavalli-Sforza 1984) only described its spread within Europe. Liu et al. (2019) have since published a review of the globalisation of wheat and rye crops, using more recent archaeological findings and including Asia, Africa and the Americas. Finding patterns and relationships between risk-alleles and CD prevalence around the world would benefit from the updated information. Therefore, this investigation used the most recent data available and an ancestry-based approach to attempt to find patterns of association between CD risk-allele frequency and CD prevalence, as well as CD risk-allele frequency and history of agriculture of gluten-containing cereals, to propose potential evolutionary mechanisms for the genes associated with CD.

Research question

Do patterns of CD risk-allele frequency across genetic ancestral regions suggest: a) significant evidence of natural selection due to the introduction of gluten-containing cereals in the diet, and b) correlation with the prevalence of CD?

Hypothesis

There is positive selection for CD risk-alleles due to the introduction of gluten-containing cereals in the diet, reflected in increased prevalence of CD in regions with a longer history of such a diet.

Methodology

Computer programs were written in R and Python languages, and are available at: https://github.com/angenibai/snp-population-frequencies

CD Prevalence

The CD prevalence in different populations around the world were collected from the two most recent worldwide analyses (Lionetti and Catassi 2014; Singh et al. 2018). Only studies using serological[14] and biopsy[6] confirmed diagnosis were included. Studies using blood donors were not included because their health was not representative of the overall population (see Appendix Table 2 and Appendix Fig. 1).

Allele Prevalence

The alleles corresponding to the three major CD-predisposing *HLA* haplotypes were searched in the online Allele Frequencies Net Database (AFND) (González-Galarza et al. 2015). Their frequency was collected for each available population, recording ethnicity of population if available (see Appendix Table 3 and Appendix Fig. 2).

A list of SNPs[15] associated with the phenotype[12] 'CELIAC DISEASE' was collected from the online Ensembl 97 database (Zerbino et al. 2017). Duplicate SNP IDs, SNPs without an associated risk-allele[13], and SNPs without tagged associated risk-alleles in the 1000 Genomes Project (1KGP) database (Clarke et al. 2016) were filtered out (see Appendix Table 4).

Using the SNPediaR library in R, a list of 8000 random SNP IDs was collected from SNPedia (Cariaso & Lennon 2011). Then, Python code was written to access the chromosome and location of each SNP in the Ensembl 97 database to select 30 random SNPs and their ancestral alleles from each of the 22 autosomes[5], ensuring no two SNPs were within 3000 base pairs of one another to create a representation of the genome as a whole.

Categorisation by Ancestry

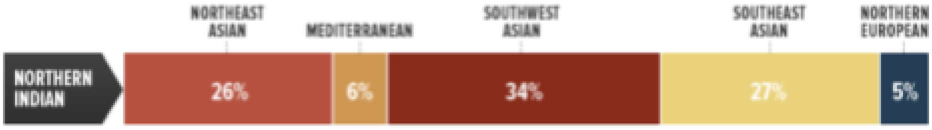
Each population from which data was collected needed to be categorised by their genetic ancestry. Referencing The Genographic Project (Behar et al. 2007), the ancestral regions[4] were identified as

European, Mediterranean, Native American, Northeast Asian, Northern European, Southeast Asian, Southwest Asian and Sub-Saharan African.

Duration of gluten-containing cereal consumption in these ancestral regions was represented by categorising the regions using the history of the spread of agriculture for two major gluten-containing cereals, wheat and rye, detailed by Liu et al. (2019). Appendix Table 5 indicates the categories for wheat and rye agriculture (WRA) duration.

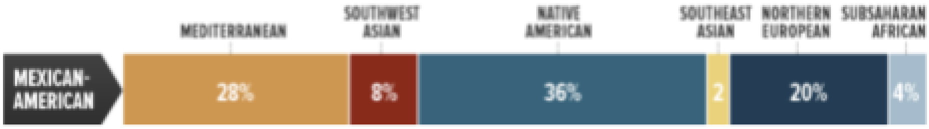
For populations identified by ethnicity, reference populations in The Genographic Project were used to categorise each ethnic group by its majority (> 50%) ancestral region. If there wasn't a majority ancestral region but regions from a single WRA category made a majority, the ethnic group was categorised by the ancestral region of greatest percentage. Otherwise, the data was excluded in order to minimise genetic heterogeneity within regions. Examples of this are shown in Figs. 3 and 4.

Fig. 3: Categorising the Northern Indian reference population



No single region makes up a majority. NE Asian and SE Asian from WRA category 3 make up majority with 53%. Population is categorised as SE Asian because it makes up the larger percentage compared to NE Asian.

Fig. 4: Categorising Mexican-American reference population



No single region forms a majority. The greatest percentage from a single WRA category is 48% from Mediterranean and N European, which is still a minority. Hence, data from this population was excluded.

For populations identified by country, the CIA World Factbook (*The World Factbook* 2018) was used to determine ethnic groups within each country. Each population was categorised into the ancestral region to which the majority of its country's ethnic groups belonged using the method outlined in the previous paragraph. See Appendix Table 6 for a full categorisation.

Pooled values for CD prevalence and haplotype frequency were calculated by ancestral region in Excel (Appendix Tables 7 and 8). A Python program was written to access the associated allele frequency in each population in the 1KGP[1] database for each of the collected SNPs, pooling the frequencies by ancestral region. See Table 2 for WRA[16] category and CD prevalence by ancestral region.

Analysis Methods

All statistical analyses (Table 1) used a confidence interval of 95% (alpha=0.05) and all linear regression analyses used the least-squares method. Using linear regression created the best fitting line modelling the relationship between the independent and dependent variables. The coefficient revealed the direction of the association relationship, the R^2 indicated the percentage of variation in the data that could be explained by the independent variable (degree of association), and the p-value suggested the significance of the association.

Table 1: Summary of regression analyses conducted

Investigating	**Independent variable**	**Dependent variable**	**Method**
Effect of WRA duration on CD prevalence in ancestral regions	WRA category	CD prevalence	In Excel. Results in Table 3 and Fig. 5
Effect of the frequency of an allele on CD prevalence	Allele frequency	CD prevalence	Python program written to conduct the regression for each of the collected haplotype and SNPs.
Effect of duration of WRA on the frequency of an allele	WRA category	Allele frequency	

The percentage of SNPs that achieved significance in the regression was compared between the random set of SNPs, the non-*HLA* risk-alleles, and the *HLA* risk-haplotypes (Table 4). The results for the random set provided a control that could indicate the extent of association due to natural genetic variation processes.

It was expected that about 5% of the random SNPs would achieve significance by chance, due to the 95% confidence interval used. A higher percentage of random SNPs achieving significance in both regressions implied that natural genetic variation was associated with the variables. It was necessary to conduct further analysis to find risk-alleles that were significantly more strongly associated than what could be expected from unrelated alleles.

Another program was written in Python to compare regression results between the risk and non-risk datasets. For each risk-allele and -haplotype, its R^2 was compared to the R^2 values of the random SNPs, counting the number of random SNPs for which it displayed a stronger association. This was then expressed as a percentage of the total number of SNPs in the randomly selected dataset. A result of over 95% indicated an association more significant than could be explained by neutral genetic variation.

The functions of the risk-alleles which achieved 95% or above were researched to evaluate potential reasons for selection. There was also the possibility that the risk-allele itself didn't have a function, but had been inherited in conjunction with a functional variant experiencing selection, in a phenomenon known as linkage disequilibrium (Slatkin 2008). Tools available on the Ensembl 97 browser and the European Bioinformatics Institute (EBI) website (EMBL-EBI 2018) were used to investigate the presence of linkage disequilibrium between variants. If a linked functional variant was found, the same process of regression and statistical analysis was conducted for the new variant.

This investigation did not involve any live humans or animals. All data related to humans was collected from existing sources published under informed consent.

Results

The groupings by ancestral region are listed in Table 2, along with their category corresponding to their duration of wheat and rye agriculture (WRA), and prevalence of CD within the region. The results of the regression analysis investigating the correlation between WRA and CD prevalence are presented in Table 3 and graphed in Figure 5.

Table 2: WRA duration and CD prevalence by ancestral region

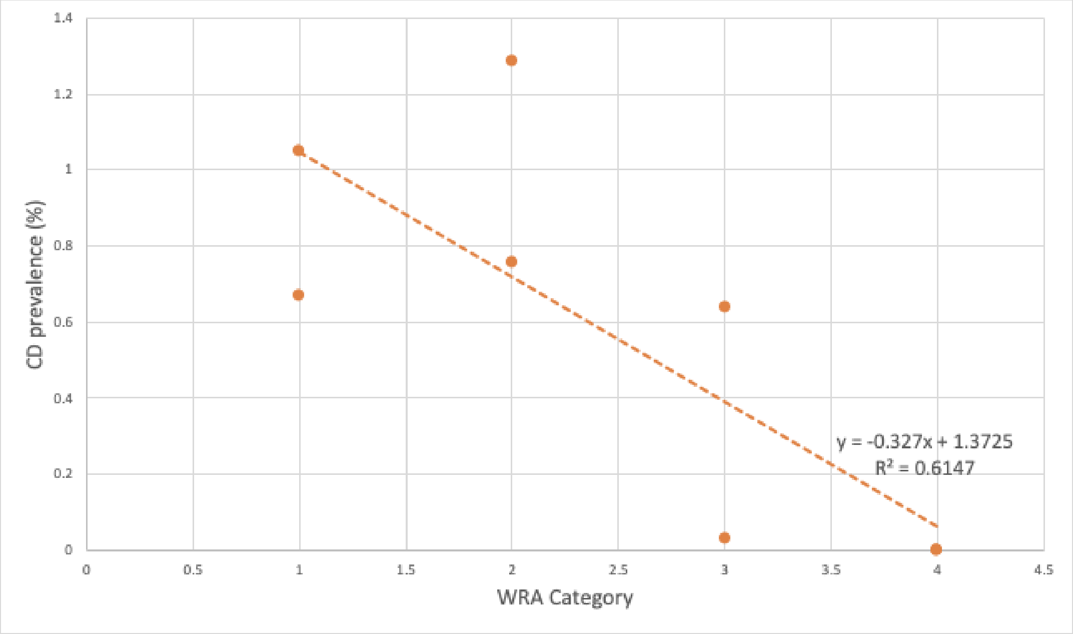
Ancestral region	**Wheat and rye agriculture duration***	**CD prevalence (%)**
European	1	0.67
Mediterranean	1	1.05
Native American	4	0.00
Northeast Asian	3	0.03
Northern European	2	1.29
Southeast Asian	3	0.64
Southwest Asian	2	0.76
Sub-Saharan African	4	0.00

* 1: Pre 5000 BCE, 2: 5000 to 2500 BCE, 3: 2500 BCE to 0 AD, 4: Post 0 AD

Table 3: Regression analysis of WRA category and CD prevalence by ancestral region

Coefficient	R^2	p-value
-0.327	0.614746464	0.021273034

Figure 5: Effect of duration of WRA on CD prevalence



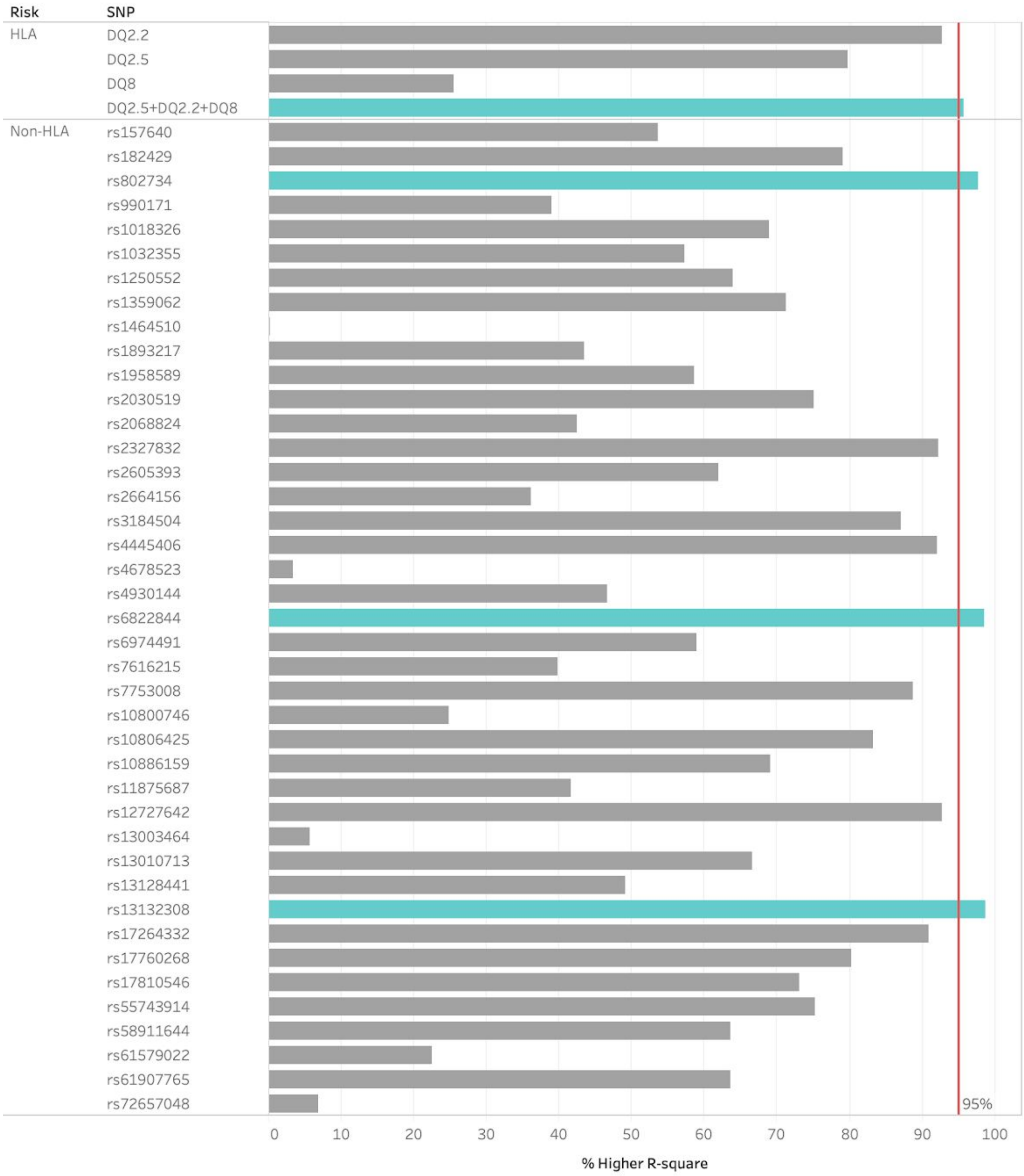
After the regression analyses outlined in Table 1 were carried out, the percentage of alleles that reached statistical significance are presented in Table 4. The alleles are separated into random SNPs, non-HLA risk-alleles[13] and risk-haplotypes[13]. Of the 660 random SNPs initially selected, data was unavailable for eight SNPs.

Table 4: Percentage of alleles which reached significance (p-value < 0.05) in regression analyses

Regression	Random SNP Sample	Non-HLA CD Risk-Alleles	HLA CD Risk-Haplotypes
CD prevalence against allele frequency	153/652 = 23.5%	12/41 = 29.3%	2/3 = 66.7%
Allele frequency against wheat agriculture history	80/652 = 12.3%	9/41 = 22.0%	0/3 = 0%

Figures 6 and 7 graph each risk-allele by the percentage of random SNPs for which the association of its frequency with CD prevalence and WRA[16] category was higher, eg. in the regression of allele frequency and CD prevalence (Fig. 6), the R^2 value of haplotype DQ2.2 was higher than approximately 92% of the R^2 values of SNPs from the randomly selected dataset. Alleles that exceeded 95% are presented in Tables 5 and 6 with further details.

Figure 6: CD risk-alleles by proportion of random SNPs for which the association with CD prevalence was higher



% Higher R-square as an attribute for each SNP broken down by Risk. Color shows details about 95% threshold met.

95% threshold met

🟦 Yes

⬛ No

Figure 7: CD risk-alleles by proportion of random SNPs for which the association with WRA category is higher

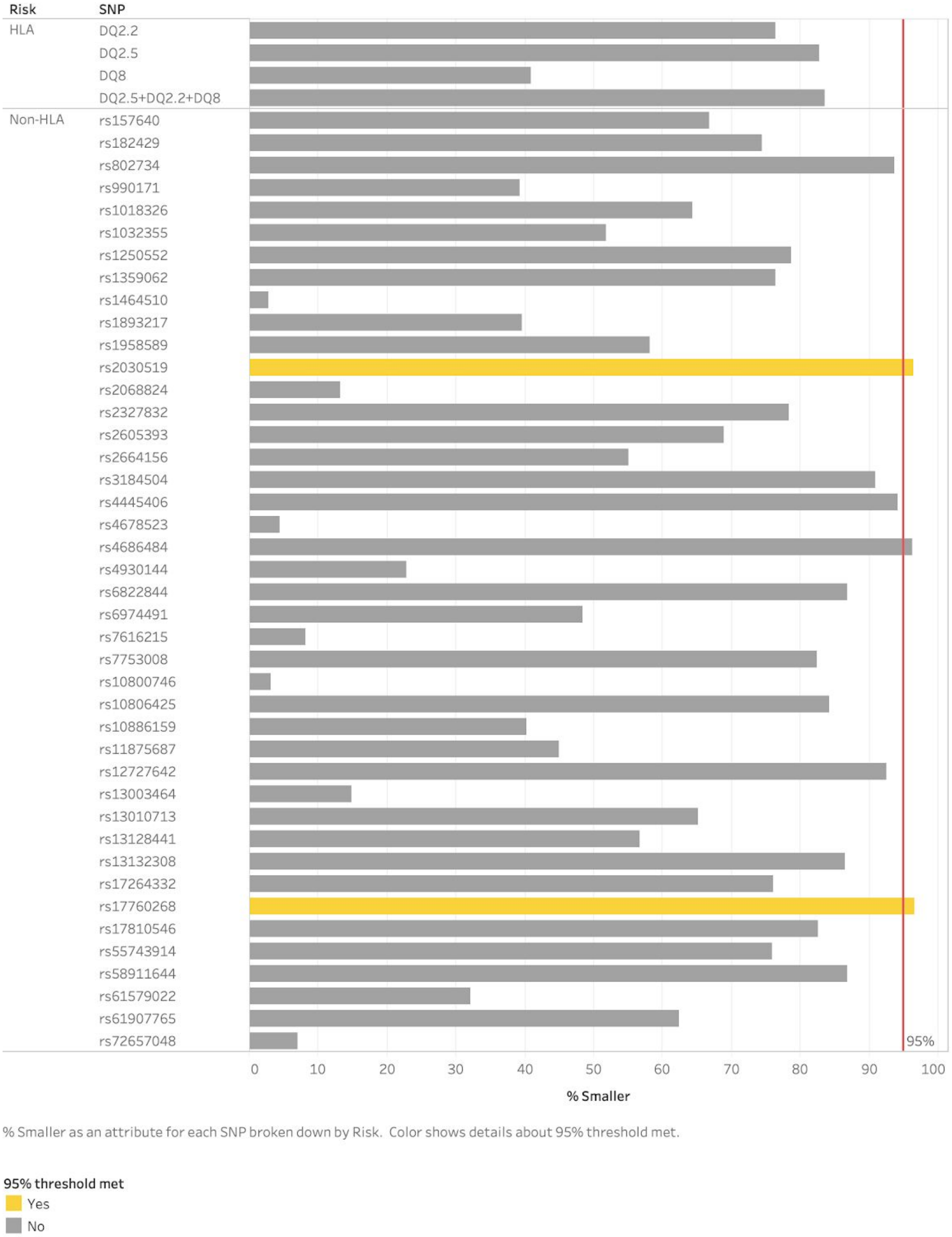


Table 5: Risk-alleles more strongly associated with CD prevalence than 95% of random alleles

ID	Risk-allele	Chromosome	Coefficient	R^2	p-value
Summed HLA risk-haplotypes	DQA1*0501/ DQB1*0201, DQA1*0201/ DQB1*0202, DQA1*03/DQB1*0302	6	6.546606846	0.831556799	0.011301183
rs802734	G	6	7.256437076	0.876541932	0.000617474
rs6822844	G	4	-8.30111685	0.914768141	0.000200009
rs13132308	A	4	-8.379453416	0.93175367	6.4e-05

Table 6: Risk-alleles more strongly associated with duration of WRA than 95% of random alleles

ID	Risk-allele	Chromosome	Coefficient	R^2	p-value
rs2030519	A	3	0.072847009	0.729038339	0.006975628
rs17760268	C	17	0.030437551	0.755376943	0.005068555

After SNP rs4686484 was found to be in linkage disequilibrium[10] with rs2030519, regression analysis was conducted of WRA duration and frequency of its risk-allele. The R^2 was used to calculate the number and percentage of random alleles for which it had a stronger association in comparison. These results are presented in Table 7.

Table 7: Regression analysis of duration of WRA and rs4686484 (G) allele frequency compared with random alleles

ID	Coefficient	R^2	p-value	Weaker random associations (n)	Weaker random associations (%)
rs4686484	-0.072813723	0.718900283	0.007826865	627	96.16564417

Discussion

Effect of duration of wheat and rye agriculture (WRA) on CD prevalence

The linear regression of duration of WRA and CD prevalence is graphed in Fig. 5, with results in Table 3 showing statistical significance ($p = 0.0213$). The R^2 value (shown in Table 3) predicts that 61% of the variance in CD prevalence between ancestral regions can be attributed to their duration of WRA. The negative coefficient indicates that a **longer** history of WRA is linked with a **higher** prevalence of CD. This relationship is counterintuitive to natural selection, because one would expect that genes which predispose to a condition that reduces fitness would be selected against, and the condition would decrease in populations over time. However, this does not appear to happen in this context and the result instead agrees with the 'CD evolutionary paradox' (Morrell and Melby 2017).

Significance of relationships

The correlation between allele frequencies and CD prevalence, and allele frequencies and wheat and rye agriculture (WRA) history between ancestral regions was modelled using a control group of randomly selected SNPs[15] from across the human genome. Six risk-haplotypes and -SNPs[13] were found to have a **stronger** association with either CD prevalence or WRA history when compared to 95% of the random SNP dataset (Figures 6 and 7) and summarised in Tables 5 and 6. The p-value for these correlations were each $p < 0.05$, which revealed a statistically significant association for these six alleles that is greater than can be attributed to neutral genetic processes (random genetic drift). This means that these six alleles appear to have been favoured by natural selection. The following discussion evaluates the capacity for these associations to be explained by mechanisms of selection.

Effect of risk-allele frequency on CD prevalence

Linear regression analysis using independent variable allele frequency and dependent variable CD prevalence aimed to test the significance of the CD risk-alleles in corresponding to actual incidence of CD. As shown in Table 5, the summed *HLA* haplotypes and three SNPs from the CD background risk network - rs802734, rs6822844 and rs13132308 - were found to be significantly associated with CD prevalence.

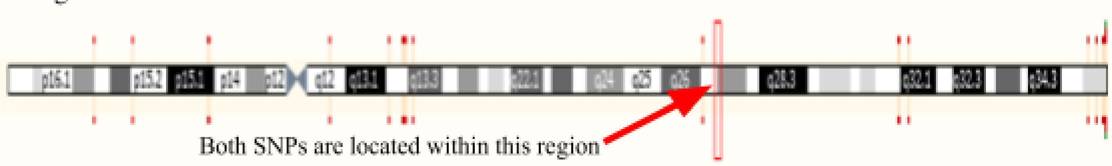
The positive coefficients of summed ***HLA* haplotypes** and **rs802734** indicate association between **higher** allele frequency and **higher** CD prevalence between ancestral regions.
Notably, significance was not reached for any of the individual *HLA* CD-predisposing haplotypes (DQ2.5, DQ2.2 or DQ8). This suggests the comparatively weaker effect of individual alleles on the development of CD. Only when their frequencies were summed to represent *HLA* CD-predisposing haplotypes as a whole did they have significance in contributing to CD prevalence across populations.

Both the *HLA* gene and rs802734 are located on chromosome 6, but rs802734 is a non-coding intergenic variant. This means that it is located between the sections of DNA that code for genes. Bondar et al. (2014) suggested it may influence the expression of the neighbouring *THEMIS* and *PTPRK* genes. *THEMIS* codes for a protein with a regulatory role in T-cells, which can control abnormal immune responses to gluten in CD patients. Hence, its association to this gene suggests it may play a role in predisposition to CD across populations.

However, SNPs **rs6822844** and **rs13132308** have negative coefficients, with **higher** allele frequency associated with a **lower** CD prevalence. This relationship appears counterintuitive.

SNP rs13132308 is another non-coding variant[11], while rs6822844 is a variant on a promoter region for the *IL2-IL21* gene. A promoter is able to initiate or prevent the transcription of its corresponding gene. *IL2* and *IL21* are interleukins, which are proteins that play a role in the immune response. Therefore, rs6822844 appears to be linked to the expression of the immune response. The SNPs exhibit strong linkage disequilibrium in the majority of populations in the 1KGP dataset (Appendix Table 9), which suggests that apparent selection for non-coding rs13132308 is due to its relation to promoter rs6822844. Their close proximity is illustrated in Fig. 8.

Fig. 8: Location of rs13132308 and rs6822844 on chromosome 4



Both SNPs are located within this region

Source: Ensembl 97

Association of rs6822844 with CD was discovered within European populations (van Heel et al. 2007). However, Maiti et al. (2010) failed to replicate this association in Argentinian subjects, and instead found significant associations with type 1 diabetes in Colombian subjects. Since rs6822844 as a CD risk-allele does not apply to all populations across the ancestral regions used in this investigation, this suggests the **significance of its association is either spurious, or caused by different selective factors**. Due to its linkage disequilibrium with rs13132308, this conclusion applies for both SNPs.

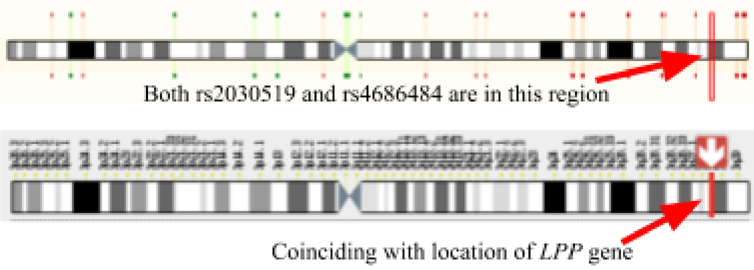
Effect of duration of gluten-containing cereal agriculture on risk-allele frequency

The linear regression analysis using independent variable wheat and rye agriculture (WRA) category and dependent variable allele frequency aimed to test the role of a region's adoption of gluten-containing cereals in the diet as a potential selective mechanism. As shown in Table 6, two CD risk-alleles, rs2030519 and rs17760268, were found have frequencies significantly associated with wheat and rye agriculture (WRA) category.

Both SNPs achieved positive coefficients in the regression, indicating an association between a **more recent** adoption of agriculture of gluten-containing cereals and **higher** allele frequency. Both SNPs are non-coding variants, meaning that they do not specifically play a role in providing instructions for the formation of proteins, but may be involved in cell function and gene expression (Zerbino et al. 2017).

The first risk-allele, **rs17760268,** has only been cited in the literature in the original study which identified it as a CD risk-variant, and no relevant SNPs were found to be in linkage disequilibrium[10] with this allele. The location of the SNP is between the genes *ANKFN1* (mainly expressed in female reproductive organ) and *NOG* (mainly expressed in the brain). Neither genes appear related to CD, which manifests in the small intestine. Without a plausible cause and effect explanation, **the association of rs17760268 appears to be spurious**.

Fig. 9 : locations of rs2030519, rs4686484 and *LNN* on chromosome 4


Both rs2030519 and rs4686484 are in this region
Coinciding with location of *LPP* gene

Sources: Ensembl 97, Wikimedia Commons

The second risk-allele, **rs2030519**, is linked with the functional variant **rs4686484** (Almeida et al. 2013), with both located on the *LPP* gene as seen in Fig. 9. This linkage disequilibrium relationship was confirmed in all 26 populations in the 1KGP dataset (Appendix Table 10). *LPP* is involved in cell mobility and cell-cell adhesion, which maintains the integrity of the tissue that lines the small intestine (Petit, Meulemans and Van de Ven 2002). Petit et al. (2002) also found that CD patients had a significantly lower expression of *LPP* gene compared to control groups. It is therefore plausible that *LPP* may play a role in CD, and may be the underlying reason for both the original identification of rs2030519 as a CD risk factor and the observed association of higher rs2030519 risk-allele frequency with regions with a longer history of wheat and rye agriculture (WRA).

This was confirmed as the regression of its linked variant rs4686484 (risk-allele G) frequency against WRA category (Table 7) showed that duration of WRA has a statistically significant effect on its frequency that is greater than can be attributed to neutral genetic processes. A negative coefficient was produced, indicating an association between **earlier** adoption of agriculture of gluten-containing cereals and **higher** allele frequency. This matches the relationship Lionetti and Catassi (2014) found between duration of wheat consumption and haplotype DQ2.

Reasons for results

In summary, the allele frequencies of rs4686484, rs802734 and the summed *HLA* haplotypes are likely to provide information that can be used in conjunction with confirmation of the CD evolutionary paradox to answer the research question: Do patterns of CD risk-allele frequency across genetic ancestral regions suggest: a) significant evidence of natural selection due to the introduction of gluten-containing cereals in the diet, and b) correlation with the prevalence of CD? For **rs4686484**, which may be involved with CD, **higher** frequency of its risk-allele is found in regions with a **longer** history of wheat and rye agriculture (WRA). Across ancestral regions, confirmed CD risk-alleles ***HLA* haplotypes** and **rs802734** display association of **higher** allele frequency corresponding with **higher** CD prevalence.

Notably, no single allele or haplotype displayed association with both CD prevalence and WRA duration. This indicates that a simple cause and effect relationship cannot explain the observations.

Higher frequency of ***HLA* risk haplotypes** in an ancestral region relates to an increased prevalence of CD. Yet, haplotype frequency isn’t significantly associated with WRA duration. This suggests that the advent of gluten-containing cereal agriculture did not have a selective effect on the CD risk-haplotypes. It also weakens the hypothesised explanation for the CD paradox which proposes that positive selection occurred for risk-alleles which protected against other conditions and pathogens (Zhernakova et al. 2010; Lionetti and Catassi 2014; Sams and Hawks 2014))

However, there is still a correlation between **higher** CD prevalence and a **longer** history of gluten-containing cereal agriculture. A possible explanation is that CD prevalence is actually more greatly influenced by recent environmental changes. Lionetti and Catassi (2014) discovered a significant relationship between a **longer duration** of wheat consumption and **greater current** wheat supply per capita between countries. This suggests that current wheat consumption in a region has a greater effect on CD prevalence than genetic factors, and is the underlying reason behind the relationship found by Lionetti and Catassi (2014) between wheat consumption history and haplotype frequency across countries. This explains why the same relationship was not detected when compared across ancestral regions within which current wheat consumption may vary. This explanation also applies to rs802734, which as part of the CD background risk network, has a lesser effect on the likelihood of developing CD than the *HLA* risk haplotypes.

Selection of SNP **rs4686484** appears influenced by the adoption of gluten-containing cereals in the diet, causing a selection pattern in the **rs2030519** CD risk-allele it is linked with. Being a functional variant on the *LPP* gene, rs4686484 is involved in maintaining the integrity of the tissue lining the small intestine, which is damaged in CD patients (Heyman et al. 2008). It can be hypothesised that as CD developed in populations alongside the adoption of gluten containing cereals in the diet, damage to the intestines was minimised for those with stronger tissue lining. Hence, rs4686484 could function as an indirect protective factor against the symptoms of CD. Since this reduces the impact of CD on reproductive fitness, the adoption of gluten-containing cereals in the diet would not have posed strong selective pressure on the CD risk-alleles. This is supported by the results from this investigation, which found that the majority of CD risk-alleles didn’t display significant signs of selection in relation to the duration of wheat and rye agriculture.

Hence, contrary to the hypothesis, there is no overall evidence of natural selection for CD risk-alleles promoted by the introduction of gluten-containing cereals in the diet of a region. CD prevalence is also not a direct reflection of the frequencies of individual risk-alleles. This highlights the inherent complexity of autoimmune diseases, such as celiac disease, which develop as a result of a combination of genetic and environmental factors.

Key limitations

The major limitations in this study stem from the availability of data.

As seen in Appendix Table 6, the number of populations within each ancestral region for a single dataset could vary greatly, with some ancestral regions only having one representative population. It would have been ideal to have multiple populations in each region from which to pool data to better represent overall genetic backgrounds and phenotypic traits. Data for the frequency of DQ2.2 wasn't available for any populations in the Native American and Southwest Asian regions. Linear regression could still be conducted because there was at least one region in each wheat and rye agriculture (WRA) category, but fewer samples limits the confidence in the analysis.

Appendix Table 2 shows that the studies for CD prevalence were conducted between 1995 and 2017. The data for background risk and non-risk alleles was published with 1KGP[1] in 2013, while data for haplotype frequency in the AFND[2] comes from studies between 1999 and 2017. The 20 year range means that the data isn't necessarily reflective of the current distribution of CD. The datasets have been used previously in peer-reviewed research (Lionetti and Catassi 2014; Singh et al. 2018), but the spread in dates does limit the power of the analysis.

SNPedia was chosen as the dataset from which to obtain the random SNP IDs because it has an easily accessible public application programming interface (API). However, the SNPs listed on the website generally have associated phenotypes[12], instead of being a mix of phenotypic and non-coding. This is likely why a higher than expected percentage of the 'random' dataset displayed a significant association with the variables, showing some level of selection bias.

Future Research

Limited information is available on the role of rs4686484 and *LPP* in relation to CD. Since this study suggests the gene may enable protection against symptoms of CD, further research into its function is required to confirm or reject this hypothesis. This could involve lab-based experiments on its coding regions of DNA.

It would also be beneficial to improve upon this study by using more powerful computational methods such as latent factor mixed models to model natural genetic variance between ancestral regions. This would allow a statistically stronger comparison of the genetic patterns of risk-alleles against the overall genome.

Conclusion

No CD risk-allele displayed an association for both CD prevalence and wheat and rye agriculture (WRA) duration between ancestral regions that was significantly stronger than that of the randomly chosen control alleles. The *HLA* haplotype and rs802734 frequencies being linked only with CD prevalence suggests their individual influence on the development of CD wasn't significant enough to experience selection as diets changed with the advent of agriculture. Increased prevalence of CD in regions of longer WRA history may instead be attributed to the selection for protective factors minimising the impact of CD on reproductive fitness and therefore reducing chance of selection against CD risk-alleles.

Reference list

Almeida, R, Ricaño-Ponce, I, Kumar, V, Deelen, P, Szperl, A, Trynka, G, Gutierrez-Achury, J, Kanterakis, A, Westra, H-J, Franke, L, Swertz, MA, Platteel, M, Bilbao, JR, Barisani, D, Greco, L, Mearin, L, Wolters, VM, Mulder, C, Mazzilli, MC, Sood, A, Cukrowska, B, Núñez, C, Pratesi, R, Withoff, S & Wijmenga, C 2013, 'Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant', *Human Molecular Genetics*, vol. 23, no. 9, pp. 2481–2489.

Ammerman, AJ & Cavalli-Sforza, LL 1984, *Neolithic transition and the genetics of populations in europe.*, Princeton University Press, Princeton.

Bondar, C, Plaza-Izurieta, L, Fernandez-Jimenez, N, Irastorza, I, Withoff, S, Wijmenga, C, Chirdo, F & Bilbao, JR 2013, 'THEMIS and PTPRK in celiac intestinal mucosa: coexpression in disease and after in vitro gliadin challenge', *European Journal of Human Genetics*, vol. 22, no. 3, pp. 358–362.

Curtis, BC 2002, 'Wheat in the world', in BC Curtis, S Rajaram & H Gómez Macpherson (eds), *Bread Wheat: Improvement and Production*, Food and Agriculture Organisation of the United Nations, Rome, viewed 29 August 2019, <http://www.fao.org/3/y4011e/y4011e04.htm>.

van Heel, DA, Franke, L, Hunt, KA, Gwilliam, R, Zhernakova, A, Inouye, M, Wapenaar, MC, Barnardo, MCNM, Bethel, G, Holmes, GKT, Feighery, C, Jewell, D, Kelleher, D, Kumar, P, Travis, S, Walters, JR, Sanders, DS, Howdle, P, Swift, J, Playford, RJ, McLaren, WM, Mearin, ML, Mulder, CJ, McManus, R, McGinnis, R, Cardon, LR, Deloukas, P & Wijmenga, C 2007, 'A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21', *Nature Genetics*, vol. 39, no. 7, pp. 827–829.

Humes, KR, Jones, NA & Ramirez, RR 2011, *Overview of Race and Hispanic Origin: 2010*, *U.S. Census Bureau*, March.

Kagnoff, MF 2007, 'Celiac disease: pathogenesis of a model immunogenetic disease', *Journal of Clinical Investigation*, vol. 117, no. 2, pp. 41–49.

Lionetti, E & Catassi, C 2014, ‘Co-localization of gluten consumption and HLA-DQ2 and -DQ8 genotypes, a clue to the history of celiac disease’, *Digestive and Liver Disease*, vol. 46, no. 12, pp. 1057–1063.

Liu, X, Jones, PJ, Motuzaite Matuzeviciute, G, Hunt, HV, Lister, DL, An, T, Przelomska, N, Kneale, CJ, Zhao, Z & Jones, MK 2019, ‘From ecological opportunism to multi-cropping: Mapping food globalisation in prehistory’, *Quaternary Science Reviews*, vol. 206, pp. 21–28.

Maiti, AK, Kim-Howard, X, Viswanathan, P, Guillén, L, Rojas-Villarraga, A, Deshmukh, H, Direskeneli, H, Saruhan-Direskeneli, G, Cañas, C, Tobön, GJ, Sawalha, AH, Cherñavsky, AC, Anaya, J-M & Nath, SK 2010, ‘Confirmation of an association between rs6822844 at theIl2-Il21region and multiple autoimmune diseases: Evidence of a general susceptibility locus’, *Arthritis & Rheumatism*, vol. 62, no. 2, pp. 323–329.

Masel, J 2011, ‘Genetic drift’, *Current Biology*, vol. 21, no. 20, pp. R837–R838.

Morrell, K & Melby, MK 2017, ‘Celiac Disease: The Evolutionary Paradox’, *International Journal of Celiac Disease*, vol. 5, no. 3.

Petit, MMR, Meulemans, SMP & Van de Ven, WJM 2002, ‘The Focal Adhesion and Nuclear Targeting Capacity of the LIM-containing Lipoma-preferred Partner (LPP) Protein’, *Journal of Biological Chemistry*, vol. 278, no. 4, pp. 2157–2168.

Sams, A & Hawks, J 2013, ‘Patterns of Population Differentiation and Natural Selection on the Celiac Disease Background Risk Network’, in V De Re (ed.), *PLoS ONE*, vol. 8, no. 7, p. e70564.

Sams, A & Hawks, J 2014, ‘Celiac Disease as a Model for the Evolution of Multifactorial Disease in Humans’, *Human Biology*, vol. 86, no. 1, pp. 19–36.

Simoons, FJ 1981, ‘Coeliac disease as a geographic problem’, in N Kretchmer & DN Walcher (eds), *Food, nutrition and evolution*, Masson, New York, pp. 179–199.

Singh, P, Arora, A, Strand, TA, Leffler, DA, Catassi, C, Green, PH, Kelly, CP, Ahuja, V & Makharia, GK 2018, ‘Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis’, *Clinical Gastroenterology and Hepatology*, vol. 16, no. 6, pp. 823-836.e2.

Slatkin, M 2008, 'Linkage disequilibrium — understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485.

Taylor, AK, Lebwohl, B & Snyder, CL 2019, *Celiac Disease*, *Nih.gov*, University of Washington, Seattle, viewed 11 June 2019, <https://www.ncbi.nlm.nih.gov/books/NBK1727/>.

Trynka, G, Hunt, KA, Bockett, NA, Romanos, J, Mistry, V, Szperl, A, Bakker, SF, Bardella, MT, Bhaw-Rosun, L, Castillejo, G, de la Concha, EG, de Almeida, RC, Dias, K-RM, van Diemen, CC, Dubois, PCA, Duerr, RH, Edkins, S, Franke, L, Fransen, K, Gutierrez, J, Heap, GAR, Hrdlickova, B, Hunt, S, Izurieta, LP, Izzo, V, Joosten, LAB, Langford, C, Mazzilli, MC, Mein, CA, Midah, V, Mitrovic, M, Mora, B, Morelli, M, Nutland, S, Núñez, C, Onengut-Gumuscu, S, Pearce, K, Platteel, M, Polanco, I, Potter, S, Ribes-Koninckx, C, Ricaño-Ponce, I, Rich, SS, Rybak, A, Santiago, JL, Senapati, S, Sood, A, Szajewska, H, Troncone, R, Varadé, J, Wallace, C, Wolters, VM, Zhernakova, A, Thelma, BK, Cukrowska, B, Urcelay, E, Bilbao, JR, Mearin, ML, Barisani, D, Barrett, JC, Plagnol, V, Deloukas, P, Wijmenga, C & van Heel, DA 2011, 'Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease', *Nature Genetics*, vol. 43, no. 12, pp. 1193–1201.

Volta, U & Villanacci, V 2011, 'Celiac disease: diagnostic criteria in progress', *Cellular & Molecular Immunology*, vol. 8, no. 2, pp. 96–102.

Zhernakova, A, Elbers, CC, Ferwerda, B, Romanos, J, Trynka, G, Dubois, PC, de Kovel, CGF, Franke, L, Oosting, M, Barisani, D, Bardella, MT, Joosten, LAB, Saavalainen, P, van Heel, DA, Catassi, C, Netea, MG & Wijmenga, C 2010, 'Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection', *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 970–977.

Databases

Behar, DM, Rosset, S, Blue-Smith, J, Balanovsky, O, Tzur, S, Comas, D, Mitchell, RJ, Quintana-Murci, L, Tyler-Smith, C & Wells, RS 2007, 'The Genographic Project Public Participation Mitochondrial DNA Database', *PLoS Genetics*, vol. 3, no. 6, p. e104.

Cariaso, M & Lennon, G 2011, 'SNPedia: a wiki supporting personal genome annotation, interpretation and analysis', *Nucleic Acids Research*, vol. 40, no. D1, pp. D1308–D1312.

Clarke, L, Fairley, S, Zheng-Bradley, X, Streeter, I, Perry, E, Lowy, E, Tassé, A-M & Flicek, P 2016, ‘The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data’, *Nucleic Acids Research*, vol. 45, no. D1, pp. D854–D859.

EMBL-EBI 2018, *The European Bioinformatics Institute*, EMBL-EBI, viewed 23 August 2019, <https://www.ebi.ac.uk/>.

González-Galarza, FF, Takeshita, LYC, Santos, EJM, Kempson, F, Maia, MHT, Silva, ALS da, Silva, ALT e, Ghattaoraya, GS, Alfirevic, A, Jones, AR & Middleton, D 2015, ‘Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations’, *Nucleic Acids Research*, vol. 43, no. D1, pp. D784–D788.

The World Factbook 2018, Central Intelligence Agency, viewed 21 June 2019, <https://www.cia.gov/library/publications/the-world-factbook/index.html>.

Zerbino, DR, Achuthan, P, Akanni, W, Amode, MR, Barrell, D, Bhai, J, Billis, K, Cummins, C, Gall, A, Girón, CG, Gil, L, Gordon, L, Haggerty, L, Haskell, E, Hourlier, T, Izuogu, OG, Janacek, SH, Juettemann, T, To, JK, Laird, MR, Lavidas, I, Liu, Z, Loveland, JE, Maurel, T, McLaren, W, Moore, B, Mudge, J, Murphy, DN, Newman, V, Nuhn, M, Ogeh, D, Ong, CK, Parker, A, Patricio, M, Riat, HS, Schuilenburg, H, Sheppard, D, Sparrow, H, Taylor, K, Thormann, A, Vullo, A, Walts, B, Zadissa, A, Frankish, A, Hunt, SE, Kostadima, M, Langridge, N, Martin, FJ, Muffato, M, Perry, E, Ruffier, M, Staines, DM, Trevanion, SJ, Aken, BL, Cunningham, F, Yates, A & Flicek, P 2017, ‘Ensembl 2018’, *Nucleic Acids Research*, vol. 46, no. D1, pp. D754–D761.

Acknowledgements

Thank you to Dr Jason Bragg (Royal Botanical Garden, Sydney) for offering guidance in computational methods of analysing genetic data and interpretation of evolutionary processes, Dr Lloyd Dawe for assisting with understanding the mathematical implications of statistical analysis, and my Science Extension teachers for their tireless support and wisdom.

Appendices

Appendix Table 1: Risk Assessment

Identify the hazard	**Strategies to minimise the hazard**	**Assessment of risk**	**What if something goes wrong?**	**Packing up**
Neck and back problems from sitting in front of a computer.	Keep laptop at eye level and maintain proper posture when sitting.	1+1 = LOW RISK	Take a longer break, do some neck stretches.	NA
Headaches from extended periods using a computer.	Take breaks every 30 minutes and stay hydrated.	1+1 = LOW RISK	Go outside to breathe fresh air for at least 30 minutes. Take painkillers if headache doesn't subside.	NA
Eye strain from staring at a computer screen.	Take breaks from looking at any screens every 30 minutes.	1+2 = MODERATE	Do eye exercises, massage eye muscles.	NA
Loss of data by spilling water or liquid onto a laptop	Back up data after every night. Keep drinks away from laptop.	1+2 = MODERATE	Switch off the laptop, remove from spill, and immediately wipe up the spill.	NA

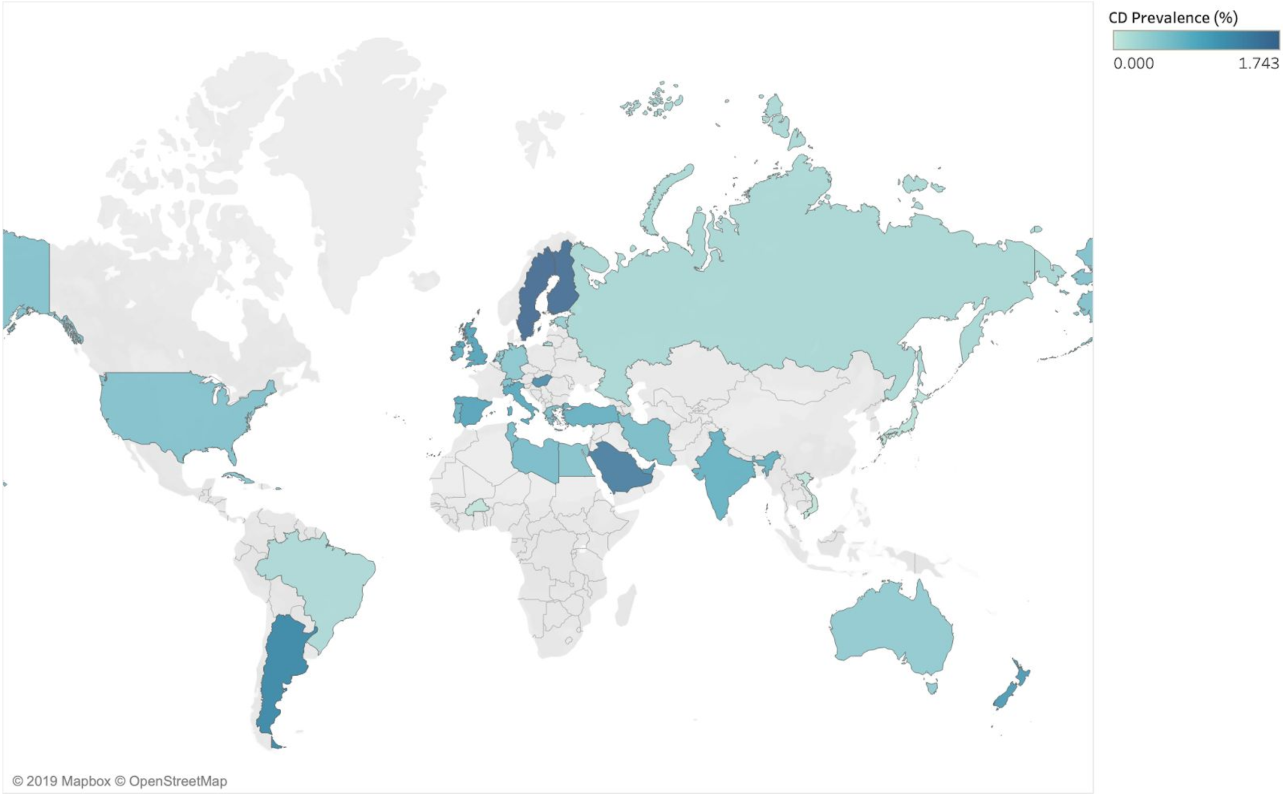
What is the potential impact or consequence?	**What is the likelihood of the event happening?**	**Assess risk**	**Action**
1 = MINOR First Aid required with little or no lost time	1 = LOW It could happen but only rarely	1 – 2 = LOW RISK	Proceed with caution
2 = MODERATE Medical treatment required, some lost time	2 = MODERATE It could occasionally happen	3 – 4 = MODERATE	Consult with teacher
3 = SERIOUS Medical treatment required, extended lost time	3 = HIGH It could frequently happen	5 – 6 = HIGH	Reassess the need to perform practical/ consult with teacher

Appendix Table 2: Prevalences of CD in different countries collected from worldwide analyses

Country	Extra Location	Age range	Prevalence	Population size	Prevalence (%)	Year of Pub
Algeria		Children 2-15 yrs	56	989	5.66	1999
Argentina		Children	28	2,219	1.26	2009
Argentine		Adults	12	2,000	0.60	2001
Australia		Adults	12	3,011	0.40	2001
Australia		Adults	14	3,011	0.46	1995
Brazil	Brasilia	Children 1-14 yrs	11	2,034	0.54	2003
Brazil	Brazilian Northeastern states of Bahia, Piaui, and Sergpipe (Sub Saharan African derived)	Adults and children	0	840	0.00	2012
Brazil	Kaingang and Guarani Indians	Adults and children	0	321	0.00	2010
Burkina Faso	Mossi (ethnic group from northern Ghana)	Adults	0	600	0.00	2002
Cuba	Nationwide	Adults and children	1	200	0.50	2007
Egypt	Cairo	Children 7mths-18yrs	8	1,500	0.53	2008
Estonia	Tartu County	Children	4	1,160	0.34	1999
Finland	Northern Finland	Children 7-16yrs	37	3,654	1.01	2003
Finland	Country wide	Adults	113	4,846	2.33	2010
Finland	Country wide	Adults	85	6,403	1.33	2010
Finland	Paijat Haime Hospital District	Elderly 52-74yrs	60	2,815	2.13	2008
Germany		Adults	8	3,098	0.26	2010
Germany	Leutkirch	Adults	8	2,157	0.37	2002
Germany	Nationwide	Children 1-17yrs	98	12,741	0.77	2015
Greece	Thessaloniki, Heraklion, and Agrinio	Children <5yrs	7	1,080	0.65	2013
Hungary	Central district	Children	5	427	1.17	1999
Hungary	Jász-Nagykun-Szolnok County	Children	37	2,690	1.38	2005
India	Punjab, north India	Children 3-17yrs	14	4,347	0.32	2006
India	North India	Children 6-12mths	4	400	1.00	2009
India	Haryana	Children and adults	31	2,879	1.08	2011
Iran		Adults	27	2,799	0.96	2006
Iran		Adults	7	1,440	0.49	2008
Iran		Children 13yrs	3	634	0.47	2012
Ireland	Northern Ireland	Adults	15	1,823	0.82	1997
Italy		Children	30	3,188	0.94	2004
Italy		Children 10-19yrs	31	2,645	1.17	2010
Italy		Adults	32	4,781	0.67	2010
Japan	Nationwide	Adults	1	2,000	0.05	2017
Libya		Children 5-17yrs	19	2,920	0.65	2011
Netherlands		Children 2-4yrs	31	6,127	0.51	1999

New Zealand		Adults	12	1,064	1.13	2000
Portugal		Children 15yrs	4	536	0.75	2006
Republic of San Marino		Adults	4	2,237	0.18	1997
Russia	Karelia region	Children 6-14yrs	4	1,988	0.20	2008
Saudi Arabia	Riyadh (capital city)	Children	119	7,930	1.50	2017
Spain	Biscay	Children 3yrs	7	830	0.84	2004
Spain	Catalonia	Children 1-14yrs	11	780	1.41	2011
Spain	Catalonia	Adults	10	3,450	0.29	2011
Spain	Catalonia (Barcelona area)	Both	21	4,230	0.50	2011
Spain	Maracena, metro district of Grenada (south)	Children	6	198	3.03	2015
Spain	Langreo (northern Spain)	Children and adults	3	1,170	0.26	2000
Spain	Madrid (central)	Children	21	3,378	0.62	2002
Sweden	Västerbotten and Norrbotten counties	Adults	10	1,894	0.53	1999
Sweden		Children 2.5yrs	9	690	1.30	2001
Sweden		Children 12yrs	212	7,274	2.91	2009
Sweden		12 yrs	329	12,632	2.60	2013
Sweden	Cities and surrounding suburbs of Umea, Norrtalje, Norrkoping, Vaxjo, and Lund	Children	195	7,567	2.58	2009
Sweden	Västerbotten and Norrbotten counties	Adults	10	1,894	0.53	1999
Switzerland		Children	8	1,450	0.55	2000
Tunisia		Children 6-12yrs	42	6,286	0.67	2007
Turkey	Erzurum	Children 6mths-17yrs	7	1,263	0.55	2006
Turkey	Nationwide	Children 6-17yrs	215	20,190	1.06	2011
UAE	Al Ain Hospital	Adults	14	1,197	1.17	2014
United Kingdom	Cambridge	Adults	85	7,550	1.13	2003
United Kingdom	Nationwide	Children 7yrs	54	5,470	0.99	2004
United Kingdom	Nationwide	Children 12-15yrs	17	1,975	0.86	2010
United Kingdom	Nationwide	Adults	69	4,656	1.48	2010
United Kingdom	Wales	Young adults	6	1,000	0.60	2004
USA		Children 6-17yrs	26	3,421	0.76	2014
USA	Nationwide	Adults	83	11,690	0.71	2014
Vietnam	Hanoi	Children 2-18yrs	0	1,961	0.00	2016

Appendix Fig. 1: Global variation of CD prevalence



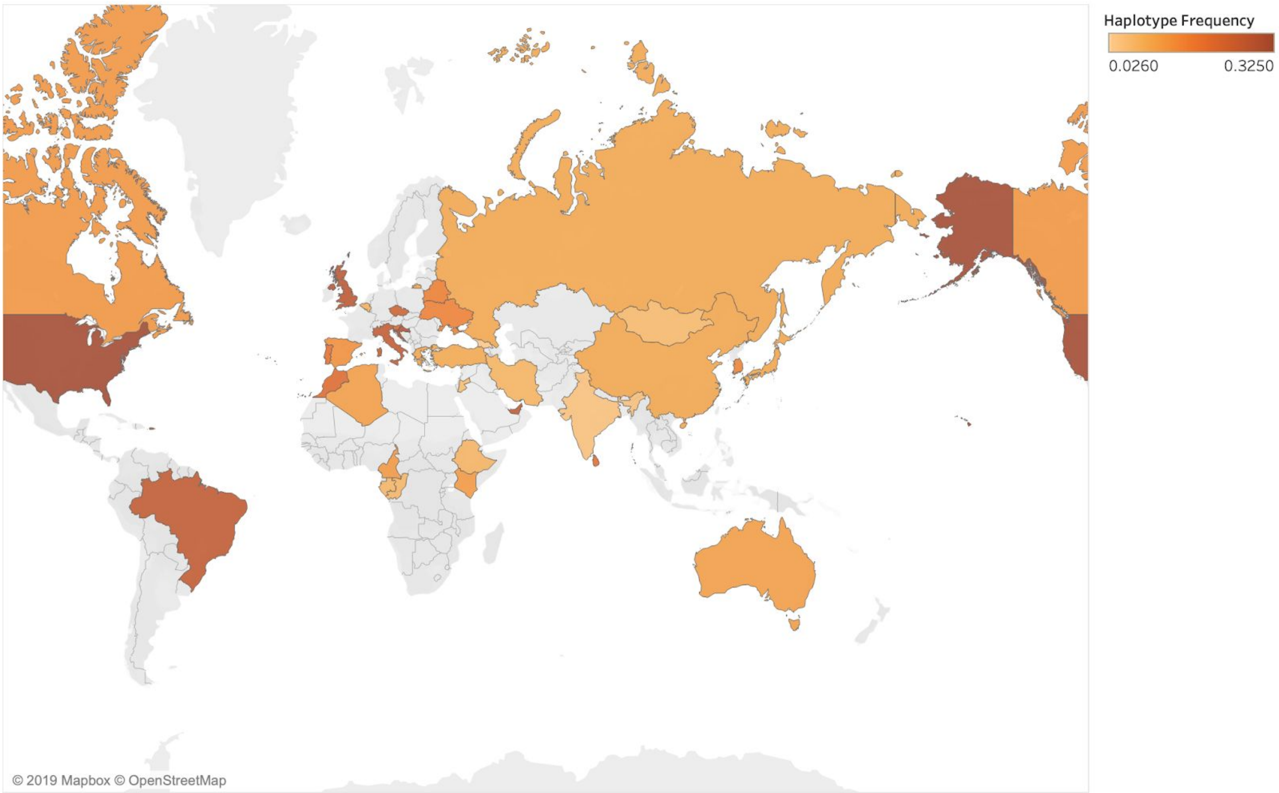
Appendix Table 3: Occurrence of HLA risk haplotypes DQ2.5, DQ2.2 and DQ8 from AFND

		DQ2.5		DQ2.2		DQ8		Summed	
Country	Ethnic Origin	Sample size	n	Sample size	n	Sample size	n	Sample size	n
Belarus	Caucasian	105	9	NA	NA	105	10	105	19
Belarus	Caucasian	100	11	NA	NA	100	8	100	19
Belarus	Caucasian	70	5	NA	NA	70	4	70	9
Belgium	Caucasian	NA	NA	715	56	NA	NA	715	56
Brazil	Caucasian	641	57	641	67	641	49	641	173
Croatia	Caucasian	63	8	63	6	63	5	63	19
Czech Republic	Caucasian	180	17	180	15	180	12	180	44
Georgia	Caucasian	80	1	NA	NA	80	1	80	2
Morocco	Caucasian	98	12	98	16	NA	NA	98	28
Slovenia	Caucasian	140	31	NA	NA	NA	NA	140	31
Turkey	Caucasian	250	24	NA	NA	NA	NA	250	24
Ukraine	Caucasian	138	7	NA	NA	138	11	138	18
Ukraine	Caucasian	102	8	NA	NA	102	12	102	19
Greece	Caucasian	246	15	NA	NA	246	9	246	25
Iran	Kurd	100	7	NA	NA			100	7
Italy	Caucasian	53	3	NA	NA	53	2	53	6
Italy	Caucasian	87	19	NA	NA	87	9	87	27
Italy	Caucasian	91	21	NA	NA	91	8	91	29
Italy	Caucasian	87	19	NA	NA	87	5	87	23
Italy	Caucasian	91	19	NA	NA	91	3	91	22
Italy	Caucasian	91	20	NA	NA	91	8	91	28
Italy	Caucasian	93	24	NA	NA	93	7	93	31
Italy	Caucasian	91	18	NA	NA	91	5	91	23
Jordan	Arab	146	4	146	3	146	1	146	8
Algeria	Arab	106	12	NA	NA	NA	NA	106	12
Morocco	Arab	98	2	98	3	98	4	98	9

Morocco	Arab	98	17	NA	NA	98	8	98	25
Portugal		130	5	130	13	130	8	130	26
Spain	Caucasian	173	10	173	11	173	3	173	24
Tunisia	Arab	100	26	100	26	100	14	100	66
UAE	Arab	52	6	52	5	52	2	52	14
Canada	Amerindian	62	2	NA	NA	62	6	62	8
China	Han	264	12	NA	NA	264	10	264	21
China	Han	59	4	NA	NA	59	2	59	7
Japan	Asian	3078	2	3078	8	3078	289	3078	299
Mongolia	Asian	85	5	NA	NA	NA	NA	85	5
Mongolia	Asian	41	3	NA	NA	NA	NA	41	3
Russia	Siberian			NA	NA	24	3	24	3
Russia	Siberian	43	1	NA	NA	NA	NA	43	1
Russia	Siberian	25	1	NA	NA	NA	NA	25	1
Russia	Siberian	68	5	NA	NA	NA	NA	68	5
Russia	Siberian	25	1	NA	NA	NA	NA	25	1
Russia	Siberian			NA	NA	17	1	17	1
Russia	Siberian	73	1	NA	NA	NA	NA	73	1
Russia	Siberian	190	3	NA	NA	NA	NA	190	3
Russia	Siberian	44	3	NA	NA	NA	NA	44	3
Russia	Siberian	22	1	NA	NA	NA	NA	22	1
South Korea	Asian	324	12	324	16	324	45	324	73
South Korea	Asian	149	3	149	8	149	7	149	17
South Korea	Asian	207	2	207	11	207	11	207	24
South Korea	Asian	467	14	467	31	467	34	467	79
England	Caucasian	177	22	NA	NA	177	27	177	50
Russia	Caucasian	81	7	NA	NA	81	12	81	19
Russia	Caucasian	126	15	NA	NA	126	12	126	27
Russia	Caucasian	202	11	NA	NA	202	9	202	20
Russia	Caucasian	200	18	NA	NA	200	18	200	36

Russia	Caucasian	156	15	NA	NA	156	10	156	24
Russia	Caucasian	121	9	NA	NA	121	14	121	23
USA	Caucasian	1899	250	1899	210	1899	180	1899	640
USA	Caucasian	220	32	220	20	220	18	220	69
Australia	Indigenous	177	20	NA	NA	NA	NA	177	20
Sri Lanka	Asian	714	27	714	67	714	51	714	145
India	Asian	190	9	NA	NA	NA	NA	190	9
India	Asian	155	7	NA	NA	155	3	155	10
India	Asian	196	7	NA	NA	196	2	196	9
India	Asian	190	6	NA	NA	NA	NA	190	6
India	Asian	188	6	NA	NA	NA	NA	188	6
India	Asian	198	7	NA	NA	NA	NA	198	7
India	Asian	202	10	NA	NA	202	4	202	14
Iran	Persian	100	6	NA	NA	NA	NA	100	6
Iran	Persian	73	8	NA	NA	NA	NA	73	8
Iran	Persian	65	4	NA	NA	NA	NA	65	4
Cameroon	African	92	6	92	6	NA	NA	92	11
Congo	African	90	6	NA	NA	NA	NA	90	6
Ethiopia	African	NA	NA	NA	NA	98	5	98	5
Ethiopia	African	NA	NA	NA	NA	83	7	83	7
Gabonese Republic	African	167	11	NA	NA	NA	NA	167	11
Kenya	African	100	9	100	4	NA	NA	100	12
China	Kazak	42	6	NA	NA	NA	NA	42	6
Mexico	Mestizo	54	2	NA	NA	NA	NA	54	2
Mexico	Mestizo	101	12	NA	NA	NA	NA	101	12
Mexico	Mestizo	160	7	NA	NA	NA	NA	160	7
Mexico	Mestizo	40	2	NA	NA	NA	NA	40	2
Nicaragua		339	13	NA	NA	NA	NA	339	13
South Africa	Mixed	159	13	NA	NA	NA	NA	159	13
USA	Mixed	496	40	NA	NA	NA	NA	496	40

Appendix Fig. 2: Global variation of summed HLA haplotype (DQ2.5+DQ2.2+DQ8) frequency



Appendix Table 4: Non-HLA CD-predisposing SNPs

SNP	Risk Allele	Chromosome
rs10800746	C	1
rs12727642	A	1
rs1359062	G	1
rs2068824	C	1
rs4445406	T	1
rs72657048	G	1
rs1018326	C	2
rs13003464	G	2
rs13010713	G	2
rs990171	A	2
rs1464510	A	3
rs17810546	G	3
rs2030519	A	3
rs2605393	G	3
rs4678523	C	3
rs61579022	A	3
rs7616215	C	3
rs1032355	C	4
rs13128441	C	4
rs13132308	A	4
rs6822844	G	4
rs10806425	A	6
rs17264332	G	6
rs182429	A	6
rs2327832	G	6
rs55743914	T	6

rs7753008	C	6
rs802734	G	6
rs6974491	A	7
rs10886159	C	10
rs1250552	A	10
rs4930144	A	11
rs61907765	T	11
rs3184504	C	12
rs1958589	C	14
rs17760268	C	17
rs11875687	C	18
rs1893217	G	18
rs2664156	C	19
rs157640	G	20
rs58911644	A	21

Appendix Table 5: Ancestral regions categorised by duration of wheat and rye agriculture

Pre 5000 BC (1)	5000 to 2500 BC (2)	2500 BC to 0 AD (3)	Post 0 AD (4)
European Mediterranean	Northern European Southwest Asian	Northeast Asian Southeast Asian	South American Sub-Saharan African

Appendix Table 6: Categorisation of populations from each dataset into ancestral regions

Ancestral region	CD prevalence dataset	AFND dataset	1000GP dataset
European	Estonia Germany Hungary Netherlands Switzerland	Belarus Caucasoid Belgium Caucasoid Brazil Caucasoid Croatia Caucasoid Czech Republic Caucasoid Georgia Caucasoid Morocco Caucasoid Slovenia Caucasoid Turkey Caucasoid Ukraine Caucasoid	Colombian in Colombia
Mediterranean	Italy Portugal San Marino Greece Libya Saudi Arabia Tunisia	Greece Caucasoid Iran Kurd Italy Caucasoid Jordan Arab Algeria Arab Morocco Arab Portugal Spain Caucasoid Tunisia Arab UAE Arab	Iberian in Spain Puerto Rican in Puerto Rico Toscani in Italy
Native American	Brazil (Kaingang and Guarani Indians)	Canada Amerindian	Peruvian in Peru
Northeast Asian	Japan Vietnam	China Asian Japan Asian Mongolia Asian Russia Siberian South Korea Asian	Chinese Dai in China Han Chinese in China Japanese in Japan Kinh in Vietnam Southern Han Chinese in China
Northern European	Finland Ireland Russia (Karelia) UK	England Caucasoid Russia Caucasoid USA Caucasoid	British in England and Scotland Finnish in Finland Northern and Western European Ancestry in

			Utah
Southeast Asian	North India	Sri Lanka Asian	Bengali in Bangladesh Sri Lankan Tamil in UK
Southwest Asian	Iran	Northeast India Asian Iran Persian	Gujarati Indian in Texas Indian Telugu in UK Punjabi in Pakistan
Sub-Saharan African	Brazil (Sub-Saharan African derived) Burkina Faso	Cameroon African Congo African Ethiopia African Gabonese Republic African Kenya African	Esan in Nigeria Gambian in Gambia Luhya in Kenya Mende in Sierra Leone Yoruba in Nigeria African Ancestry in Southwest US
Excluded	Australia Cuba Egypt New Zealand Republic of San Marino Spain USA Mixed UAE	China Kazak Mexico Mestizo Nicaragua unknown South Africa Mixed USA Mixed	African Carribean in Barbados Mexican Ancestry in California

Appendix Table 7: Prevalence of CD pooled by ancestral region

Ancestral region	CD prevalence
European	0.67
Mediterranean	1.05
Native American	0
Northeast Asian	0.03
Northern European	1.29
Southeast Asian	0.64
Southwest Asian	0.76
Sub-Saharan African	0

Appendix Table 8: Frequencies of HLA risk haplotypes pooled by ancestral region

Ancestral region	**HLA DQ2.5**	**HLA DQ2.2**	**HLA DQ8.1**	**Summed haplotypes**
European	0.096	0.095	0.076	0.267
Mediterranean	0.127	0.087	0.057	0.271
Native American	0.032	NA	0.096	0.128
Northeast Asian	0.014	0.017	0.088	0.119
Northern European	0.119	0.109	0.094	0.322
Southeast Asian	0.038	0.094	0.071	0.203
Southwest Asian	0.045	NA	0.01	0.055
Sub-Saharan African	0.07	0.047	0.068	0.185

Appendix Table 9: Pairwise disequilibrium of rs6822844 and rs13132308 in populations in 1KGP

Population	**Focus Variant**	**Variant 2**	**r^2**	**D'**
African Caribbean in Barbados	rs6822844	rs13132308	0.791304	1.000000
African Ancestry in Southwest US	rs6822844	rs13132308	0.655367	1.000000
Bengali in Bangladesh	rs6822844	rs13132308	0.855765	1.000000
Utah residents with Northern and Western European Ancestry	rs6822844	rs13132308	1.000000	1.000000
Colombian in Medellin, Colombia	rs6822844	rs13132308	1.000000	1.000000
Finnish in Finland	rs6822844	rs13132308	1.000000	1.000000
British in England and Scotland	rs6822844	rs13132308	0.968739	1.000000
Iberian populations in Spain	rs6822844	rs13132308	0.956451	1.000000
Indian Telugu in the UK	rs6822844	rs13132308	0.904405	1.000000
Kinh in Ho Chi Minh City, Vietnam	rs6822844	rs13132308	1.000000	1.000000
Mexican Ancestry in Los Angeles, California	rs6822844	rs13132308	1.000000	1.000000
Peruvian in Lima, Peru	rs6822844	rs13132308	1.000000	1.000000
Punjabi in Lahore, Pakistan	rs6822844	rs13132308	1.000000	1.000000
Puerto Rican in Puerto Rico	rs6822844	rs13132308	1.000000	1.000000
Sri Lankan Tamil in the UK	rs6822844	rs13132308	0.791837	1.000000
Toscani in Italy	rs6822844	rs13132308	0.927310	1.000000
Yoruba in Ibadan, Nigeria	rs6822844	rs13132308	0.746478	1.000000
Gujarati Indian in Houston, TX	rs6822844	rs13132308	0.734535	0.999999

$R^2 = 1$ indicates complete LD (co-inherited), $R^2 = 0$ indicates no correlation
D' = 1 indicates co-inheritance, D' = 0 indicates complete independence.

Appendix Table 10: Pairwise disequilibrium of rs2030519 and rs4686484 in populations in 1KGP

Population	Focus Variant	Variant 2	r^2	D'
African Caribbean in Barbados	rs2030519	rs4686484	1.000000	1.000000
African Ancestry in Southwest US	rs2030519	rs4686484	1.000000	1.000000
Bengali in Bangladesh	rs2030519	rs4686484	1.000000	1.000000
Chinese Dai in Xishuangbanna, China	rs2030519	rs4686484	1.000000	1.000000
Utah residents with Northern and Western European Ancestry	rs2030519	rs4686484	1.000000	1.000000
Han Chinese in Beijing, China	rs2030519	rs4686484	1.000000	1.000000
Southern Han Chinese, China	rs2030519	rs4686484	1.000000	1.000000
Colombian in Medellin, Colombia	rs2030519	rs4686484	1.000000	1.000000
Esan in Nigeria	rs2030519	rs4686484	1.000000	1.000000
Finnish in Finland	rs2030519	rs4686484	1.000000	1.000000
British in England and Scotland	rs2030519	rs4686484	1.000000	1.000000
Gujarati Indian in Houston, TX	rs2030519	rs4686484	1.000000	1.000000
Iberian populations in Spain	rs2030519	rs4686484	1.000000	1.000000
Indian Telugu in the UK	rs2030519	rs4686484	0.957710	0.999999
Japanese in Tokyo, Japan	rs2030519	rs4686484	1.000000	1.000000
Kinh in Ho Chi Minh City, Vietnam	rs2030519	rs4686484	0.974412	1.000000
Luhya in Webuye, Kenya	rs2030519	rs4686484	0.940325	1.000000
Gambian in Western Division, The Gambia	rs2030519	rs4686484	1.000000	1.000000
Mende in Sierra Leone	rs2030519	rs4686484	1.000000	1.000000
Mexican Ancestry in Los Angeles, California	rs2030519	rs4686484	1.000000	1.000000
Peruvian in Lima, Peru	rs2030519	rs4686484	1.000000	1.000000
Punjabi in Lahore, Pakistan	rs2030519	rs4686484	0.977256	1.000000
Puerto Rican in Puerto Rico	rs2030519	rs4686484	1.000000	1.000000
Sri Lankan Tamil in the UK	rs2030519	rs4686484	1.000000	1.000000
Toscani in Italy	rs2030519	rs4686484	1.000000	1.000000
Yoruba in Ibadan, Nigeria	rs2030519	rs4686484	1.000000	1.000000

R^2 = 1 indicates complete LD (co-inherited), R^2 = 0 indicates no correlation
D' = 1 indicates co-inheritance, D' = 0 indicates complete independence.